Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 121-127 recibido 06-02-08, aceptado 03-03-08

Categorización de textos biomédicos usando UMLS*

Biomedical text categorization using UMLS

José Manuel Perea Ortega
Arturo Montejo Ráez
María Teresa Martín Valdivia
Manuel Carlos Díaz Galiano

Universidad de Jaén, Campus Las Lagunillas
Edificio A3. E-23071
{jmperea,maite,amontejo,mcdiaz}@ujaen.es

Resumen: En este artículo se presenta un sistema automático de categorización de texto multi-etiqueta que hace uso del metatesauro UMLS (Unified Medical Language System). El sistema ha sido probado sobre un corpus biomédico que incluye textos muy cortos pertenecientes a expedientes de niños con enfermedades respiratorias. El corpus ha sido enriquecido utilizando las ontologías que incluye UMLS y los resultados obtenidos demuestran que la expansión de términos realizada mejora notablemente al sistema de categorización tradicional.
Palabras clave: Categorización de texto, Ontologías, UMLS, Integración de conocimiento, Expansión de términos

Abstract: In this paper we present an automatic system for multi-label text categorization which makes use of UMLS (Unified Medical Language System). Our approach has been tested on a biomedical corpus which includes very short texts belonging to expedients of children with respiratory disseases. The corpus has been enriched by using those ontologies integrated in UMLS and the results obtained show that the term expansion approach proposed greatly improves the traditional categorization system.
Keywords: Text categorization, Ontology, UMLS, Knowledge integration, Term expansion

1. *Introducción*

No cabe duda que la información es uno de los recursos fundamentales en cualquier ámbito profesional o personal. Sin embargo, en los últimos años, la cantidad de información generada diariamente de manera electrónica está creciendo de forma exponencial. De hecho, el acceso a dicha información se está convirtiendo en un gran problema. Esta saturación de información está provocando que gran parte de la investigación en nuevas tecnologías esté siendo orientada a la recuperación y uso eficiente de dicha información. Parte de esta investigación hace uso de técnicas y herramientas propias del Procesamiento del Lenguaje Natural (PLN). El PLN es una disciplina que ha demostrado a lo largo de los años que es imprescindible para mejorar la precisión de los sistemas de información (Mitkov, 2003) tales como sistemas de categorización de documentos, sistemas de recuperación de información monolingüe y multilingüe, sistemas de extracción de conocimiento, sistemas de generación automática de resúmenes...

En este trabajo se presenta un sistema de categorización de textos multi-etiqueta que ha sido entrenado en un entorno biomédico. La categorización de textos es una de las tareas fundamentales del PLN y que mas ampliamente han sido estudiadas (Sebastiani, 2002). La categorización consiste en determinar si un documento dado pertenece a un conjunto de categorías predeterminadas.

Por otra parte, una de las técnicas que han sido utilizadas para aumentar la precisión de los sistemas consiste en la integración de recursos externos que permitan obtener una información de mayor calidad. Así por ejemplo,

* Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología a través del proyecto TIMOM (TIN2006-15265-C06-03).

la integración de conocimiento mediante el uso de ontologías ha conseguido muy buenos resultados en numerosos sistemas. Por ejemplo, WordNet[1] (Miller, G.A. et al., 1993) ha sido utilizada con éxito en multitud de trabajos relacionados con recuperación de información, desambiguación e incluso categorización de textos (Martín Valdivia, Ureña López, y García Vega, 2007).

Por otra parte, en el entorno biomédico se están desarrollando muchos sistemas de información que hacen uso de recursos externos como ontologías. Los trabajos realizados demuestran que la integración de conocimiento puede ayudar a mejorar los sistemas. Por ejemplo, la ontología GO[2] (Gene Ontology) ha constituido una fuente de información incalculable para muchos investigadores que trabajan con temas relacionados con el genoma humano (Bontempi, 2007). La ontología MeSH (Medical Subject Headings) ha sido aplicada con éxito para expandir términos de las consultas en sistemas de recuperación de información (Díaz Galiano et al., 2007). Sin embargo, la mayoría de los trabajos que integran información a partir de ontologías han estado orientados a la recuperación y extracción de información más que a la categorización de texto. En un trabajo anterior (Martín Valdivia et al., 2007) hicimos uso de la ontología MeSH pero los resultados obtenidos no fueron muy prometedores. El sistema desarrollado realizaba una expansión de términos que tenía en cuenta la jerarquía de conceptos de MeSH usando los nodos padres, hijos y/o hermanos. En este artículo se propone usar el metatesauro UMLS que incluye varias ontologías médicas (entre ellas la ontología MeSH) para realizar una expansión de términos a la colección de documentos CCHMC. Con esto, se pretende conseguir una mejor categorización de textos multi-etiqueta integrando el conocimiento incluido en UMLS sobre el corpus CCHMC.

El artículo se organiza de la siguiente manera: en primer lugar, se describe brevemente la tarea de categorización de textos multi-etiqueta así como el sistema categorizador utilizado. A continuación, se presentan el corpus biomédico utilizado (el corpus CCHMC). El metatesauro UMLS se describe en la siguiente sección junto con la manera de expandir los términos del corpus. En la sección cinco se muestran los experimentos y resultados obtenidos. Finalmente, se comentan las conclusiones y trabajos futuros.

[1] http://wordnet.princeton.edu

[2] http://www.geneontology.org

2. Categorización de textos multi-etiqueta

La asignación automática de palabras clave a los documentos abre nuevas posibilidades en la exploración documental (Montejo Ráez y Steinberger, 2004), y su interés ha despertado a la comunidad científica en la propuesta de soluciones. La disciplina de la *Recuperación de Información* (RI), junto con las técnicas para el *Procesamiento del Lenguaje Natural* (PLN) y los algoritmos de *Aprendizaje Automático* (*Machine Learning*, ML) son el sustrato de donde emergen las tareas de *Categorización Automática de Textos* (Sebastiani, 2002). Los algoritmos de aprendizaje empleados van desde clasificadores lineales, probabilísticos y métodos de regresión (Joachims, 1998), (Friedman, Geiger, y Goldszmidt, 1997), (Lewis et al., 1996) a redes neuronales (Martín Valdivia, García Vega, y Ureña López, 2003; Li et al., 2002), pasando por técnicas de voto y boosting (Li et al., 2002; Bauer y Kohavi, 1999).

En la clasificación de documentos se distinguen tres casos: *categorización binaria*, cuando el clasificador debe devolver una de entre dos posibles categorías, *categorización multi-clase*, cuando el clasificador debe proporcionar una categoría de entre varias propuestas. Por último, tenemos el caso más complejo, la *categorización multi-etiqueta*, donde el clasificador debe determinar un número indefinido de clases de entre una amplia variedad de candidatas.

En cualquier caso, los sistemas de categorización automáticos se componen habitualmente de dos módulos principales: un procesador de documentos y un algoritmo de entrenamiento y clasificación. El primero transforma los textos a representaciones manejables por los segundos, generalmente siguiendo el modelo de espacio vectorial. El segundo aplica algoritmos de aprendizaje automático para modelizar los clasificadores.

El dominio biomédico ha sido uno de los más interesados en el desarrollo y progreso de este tipos de sistemas, al contar con una larga tradición en el uso de ontologías y vocabularios controlados para el manejo de documentos, siendo el multi-etiquetado el problema que se plantea en general.

BIOSIS categorizaba documentos a partir de un vocabulario de 15,000 términos biológicos que se podían resumir en 600 conceptos (Vieduts-Stokolo, 1987). Esta clasificación era jerárquica, y si sólo se consideraba el nivel primario en torno al 75 % de los conceptos quedaban cubiertos por el sistema. La precisión rozaba el 65 %.

Medical Subject Headings (MeSH) es una taxonomía de conceptos médicos usados para la categorización de documentos en la base de datos MEDLINE. El sistema desarrollado por Bruno Pouliquen (Pouliquen, Delamarre, y Beux, 2002) denominado *Nomindex* es una de las primeras propuestas para la automatización de su etiquetado. Su sistema aplicaba principalmente medidas estadísticas típicas dentro del mundo de la Recuperación de Información dando como resultado un sistema más que aceptable.

Podemos citar también el trabajo de Wright et al. (Wright et al., 1999) en el desarrollo de una herramienta para el indexado de documentos en el UMLS (siglas de *Unified Medical Language System* en inglés). Este sistema hace también uso intensivo de recursos lingüísticos como el reconocimiento de componentes nominativos (*noun phrases*) o sinónimos. Una combinación de la información en el título, el resumen y el contenido permite asignar a cada concepto del tesauro MeSH.

Nuestro enfoque se ha centrado en el uso de las ontologías médicas como un recurso para la mejora de los sistemas de categorización mediante la expansión de términos en la consulta. Con respecto a trabajos anteriores (Martín Valdivia et al., 2007), hemos modificado el método de expansión, pasando de usar exclusivamente MeSH y una expansión basada en recorridos sobre la jerarquía de términos a una expansión sobre UMLS a través de la interfaz *MetaMap Transfer*[3]. El conjunto de datos utilizado no difiere, así como el sistema de categorización y evaluación: hemos aplicado la herramienta TECAT[4] sobre el corpus CCHMC (detallado más adelante) mediante una validación cruzada. Si bien los resultados eran desalentadores, consideramos que el problema debía radicar en la ontología usada así como en la forma en que ésta fue aplicada. Es por ello que estudiar un cambio de enfoque era necesario a la hora de emitir un juicio acerca de los efectos que la integración de estos recursos producen en la categorización de textos biomédicos.

[3] http://mmtx.nlm.nih.gov/index.shtml

[4] http://sinai.ujaen.es/wiki/index.php/TeCat

3. La colección CCHMC

Esta colección de 978 documentos ha sido preparada por "The Computational Medicine Center"[5]. Dicho corpus incluye registros médicos anónimos recopilados en el departamento de radiología del Hospital infantil de Cincinnati (the Cincinnati Children's Hospital Medical Center's Department of Radiology - CCHMC) (cmc, 2007).

Estos documentos son informes radiológicos que están etiquetados con códigos del ICD-9-CM (Internacional Classification of Diseases 9th Revision Clinical Modification). Se trata de un catálogo de enfermedades codificadas con un número de 3 a 5 dígitos con un punto decimal después del tercer dígito. Los códigos ICD-9-CM son un subgrupo de los códigos ICD-9. Están organizados de manera jerárquica, agrupándose varios códigos consecutivos en los niveles superiores. Estos códigos están relacionados con enfermedades del sistema respiratorio únicamente y sus valores se establecen dentro del rango de números 460 al 519[6].

Cada documento contiene dos campos de texto a partir del cual se ha construido el cuerpo a procesar: CLINICAL_HISTORY e IMPRESSION. Ambos campos son, por lo general, muy breves, veamos un ejemplo:

CLINICAL_HISTORY: `Eleven year old with ALL, bone marrow transplant on Jan. 2, now with three day history of cough.`

IMPRESSION: `1. No focal pneumonia. Likely chronic changes at the left lung base. 2. Mild anterior wedging of the thoracic vertebral bodies.`

La brevedad de contenido nos hace pensar que la expansión de términos debería contribuir a una mejora del sistema de categorización, al aumentar el número de características representativas de cada documento. El proceso seguido para dicha expansión se describe más adelante.

[5] http://www.computationalmedicine.org

[6] Se puede consultar dicho catálogo de códigos ICD-9-CM en la dirección http://www.cs.umu.se/~medinfo/ICD9/icd9cm_group8.html

4. UMLS

UMLS[7] es un repositorio de varias ontologías biomédicas desarrollado por la Biblioteca Nacional de Medicina de Estados Unidos. UMLS integra más de 2 millones de nombres para unos 900,000 conceptos procedentes de más de 60 familias de vocabularios biomédicos, así como 12 millones de relaciones entre esos conceptos (Bodenreider, 2004). UMLS es un sistema que garantiza referencias cruzadas entre más de treinta vocabularios y clasificaciones. La mayoría de estas referencias cruzadas se realizan gracias al análisis léxico de los términos, de ahí su inclusión en la categoría de sistemas léxicos de clasificación en el dominio biomédico (Ceusters et al., 1997). Algunos ejemplos de ontologías que incorpora UMLS son ICD-9-CM, ICD-10, MeSH, SNOMED CT, LOINC, MEDLINE, WHO Adverse Drug Reaction Terminology, UK Clinical Terms, RxNORM, Gene Ontology, and OMIM.

UMLS está formado por tres componentes principales:

- El **Metatesauro**. Es la base de datos núcleo de UMLS, una colección de conceptos, términos y sus relaciones. El Metatesauro está organizado por conceptos, y cada concepto tiene atributos específicos que definen su significado y lo enlazan a sus correspondientes nombres de conceptos en las distintas ontologías que conforman UMLS. También se representan numerosas relaciones entre conceptos, tales como "*es un*", "*es parte de*", "*es causado por*", etc.

- El **Lexicón Especializado**. Es una base de datos de información lexicográfica para uso en Procesamiento de Lenguaje Natural. Contiene información sobre vocabulario común, términos biomédicos, términos encontrados en *MEDLINE* y en el propio Metatesauro. Cada entrada contiene información sintáctica, morfológica y ortográfica.

- La **Red Semántica**. Es un conjunto de categorías y relaciones usadas para clasificar y relacionar las entradas en el Metatesauro. Cada concepto en el Metatesauro se asigna al menos a un tipo semántico o categoría. Existen 135 tipos semánticos definidos y 54 relaciones entre ellos.

[7] http://www.nlm.nih.gov/research/umls

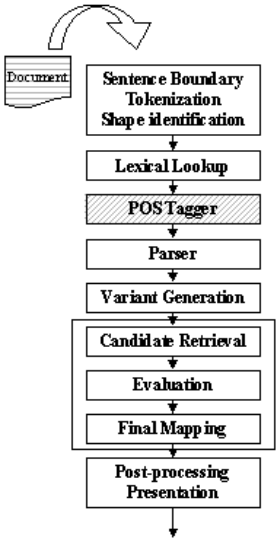

Figura 1: Procesamiento de un texto con *MetaMap*

UMLS tiene varias herramientas software de soporte como ***MetaMap***. *MetaMap* es una herramienta online que se utiliza para encontrar conceptos relevantes del Metatesauro dado un texto arbitrario. *MetaMap Transfer (MMTx)* provee la misma funcionalidad que *MetaMap* pero como programa Java. Para los experimentos de este trabajo hemos utilizado esta interfaz.

4.1. Expansión de CCHMC usando UMLS

Para expandir con UMLS cada fichero de texto de la colección CCHMC hemos utilizado la herramienta *MetaMap Transfer (MMTx)*. El texto de cada fichero se procesa a través de una serie de módulos. En primer lugar, el texto se divide en componentes como párrafos, sentencias, frases, elementos léxicos y tokens. Después, las distintas variantes se generan a partir de las frases detectadas. Los conceptos candidatos del Metatesauro UMLS son recuperados y evaluados en relación a estas frases. Los conceptos candidatos que mayor similitud tengan con la frase se organizan en un *mapping* final que será el que se utilice para la expansión de términos. Se puede observar el procesamiento que sigue el texto de un documento con *MetaMap* en la Figura 1.

El pseudocódigo seguido en los experimentos para realizar la expansión de términos a

un documento de la colección CCHMC se explica a continuación:

1. Para cada sentencia encontrada en el documento obtenemos las frases detectadas.
2. Para cada frase obtenemos su *mapping* final (mejores conceptos candidatos).
3. Para cada concepto candidato:
   - Obtenemos su nombre UMLS y lo añadimos al conjunto de términos expandidos (si no estuviera ya añadido).
   - Añadimos también al conjunto de la expansión el grupo de términos sinónimos que conforman dicho concepto, es decir, aquellos términos que aparecen en distintas ontologías de UMLS y que pertenecen al concepto en cuestión, controlando que no haya términos repetidos.

En la Figura 2 podemos ver varios ejemplos de expansión realizada con la herramienta *MetaMap Transfer (MMTx)* a un documento de la colección CCHMC, siguiendo las estrategias que se explican en el apartado 5.

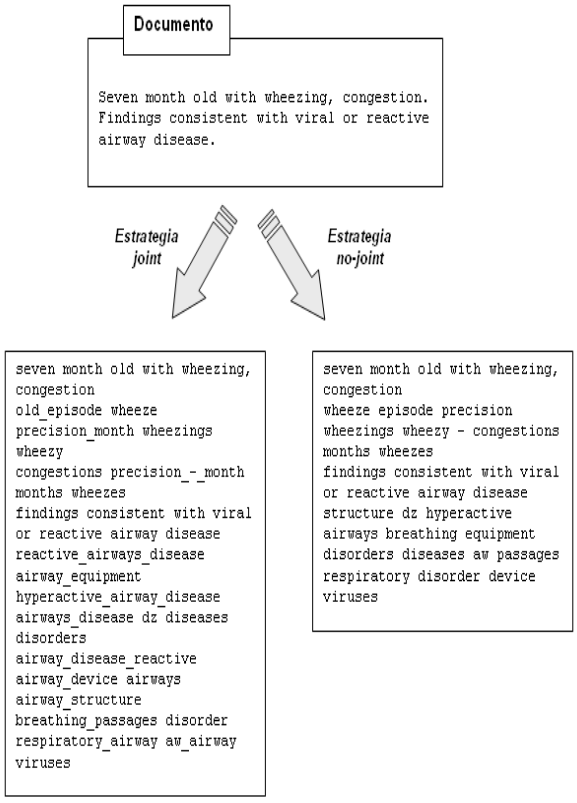

Figura 2: Ejemplos de expansión UMLS de un documento de la colección CCHMC

5. *Experimentos y resultados*

Para este trabajo se han realizado varios experimentos con distintos tipos de expansión UMLS y con diferentes algoritmos de aprendizaje automático. Concretamente se ha utilizado el algoritmo SVM (Support Vector Machine) y una red neuronal tipo perceptrón denominada PLAUM. Para estos algoritmos se han considerado sus configuraciones por defecto, sin variaciones de ningún parámetro. También se ha utilizado expansión de términos haciendo uso de una ontología médica como UMLS para incorporar información de calidad a los documentos de la colección que ayude a mejorar la categorización de los mismos. Los resultados demuestran que el uso de SVM es mejor que PLAUM cuando no se aplica expansión de términos. En cambio, PLAUM mejora cuando hemos utilizado expansión. Para todos los casos, el uso de la expansión de términos con UMLS mejora el caso base.

La expansión de los documentos de la colección CCHMC se ha realizado utilizando la ontología médica UMLS. El procedimiento seguido para realizar dicha expansión se ha descrito en el apartado 4.1. En algunas ocasiones, los términos de expansión obtenidos de la ontología estaban compuestos por más de una palabra o *token*. Esta característica nos ha permitido utilizar dos estrategias en el proceso de expansión:

- ***Estrategia joint***. Consiste en considerar los términos de expansión de más de una palabra como un único *token*. Para ello, hemos sustituido los espacios entre las palabras del término por el símbolo subrayado. De esta forma se consigue introducir más términos diferentes para el posterior proceso de clasificación.
- ***Estrategia no-joint***. Consiste en separar los *tokens* de aquellos términos de expansión formados por más de una palabra y añadirlos por separado a la expansión, comprobando que no haya *tokens* repetidos. Con esta estrategia, al contrario de lo que ocurre con la anterior, el número total de términos añadidos a los documentos de la colección es bastante inferior.

En la Figura 2 se puede observar el resulta-

	PLAUM	SVM
P	80.91 %	90.48 %
R	64.08 %	61.79 %
F1	71.52 %	**73.43 %**

Tabla 1: *Micro-averaging* sin expansión

	PLAUM	SVM
P	85.17 %	92.04 %
R	69.49 %	62.92 %
F1	**76.53 %**	74.74 %

Tabla 2: *Micro-averaging* con expansión *no-joint*

do de la aplicación de ambas estrategias de expansión a un documento de la colección.

Con respecto a la evaluación de los resultados obtenidos, las medidas consideradas son la precisión (P), la cobertura (R) y la F1, siendo ésta última la que nos da una visión más completa del comportamiento del sistema. Estas medidas han sido obtenidas mediante *micro-averaging* sobre validación cruzada en 10 particiones (*10-fold cross-validation*), es decir, repitiendo el experimento 10 veces con distintas colecciones de entrenamiento y evaluación, y calculando, cada vez, los aciertos y fallos en cada clase de forma acumulativa y calculando los valores finales sobre dichos valores acumulados. Se pueden observar los resultados obtenidos para los distintos experimentos en las tablas 1, 2 y 3 para la medida *micro-averaging*.

Si analizamos los resultados desde el punto de vista de la expansión de los documentos, se puede afirmar que la integración de UMLS mejora notablemente los resultados sin expansión. En concreto, para el algoritmo PLAUM, la medida F1 mejora en 6,54 puntos si se utiliza expansión *no-joint* y en 7,64 puntos con expansión *joint*. Para el algoritmo SVM ocurre igual pero con una diferencia más pequeña que el PLAUM (1,75 puntos con expansión *no-joint* y 3,84 puntos con expansión *joint*).

En cuanto a los algoritmos de aprendizaje utilizados, se puede observar que la expansión funciona tanto para PLAUM como para SVM, pero hay que señalar que SVM funciona mejor que PLAUM cuando no se aplica expansión de términos (2,6 puntos mejor). En cambio, con PLAUM se han obtenido mejores resultados que con SVM cuando hemos utilizado expansión de términos UMLS, aunque

	PLAUM	SVM
P	84.97 %	92.98 %
R	71.13 %	64.80 %
F1	**77.44 %**	76.37 %

Tabla 3: *Micro-averaging* con expansión *joint*

las diferencias no son muy importantes (2,33 puntos para la estrategia *no-joint* y 1,38 puntos para la expansión *joint*).

6. Conclusiones y trabajo futuro

En este trabajo se ha presentado un estudio sobre la integración de conocimiento médico en la categorización multi-etiqueta de documentos biomédicos. Para ello, se ha expandido el corpus utilizado (CCHMC) en el proceso de categorización multi-etiqueta con el tesauro médico UMLS. Para realizar el estudio se han utilizado dos algoritmos de aprendizaje como SVM y PLAUM. Aunque las diferencias encontradas entre ambos algoritmos no son determinantes, parece que PLAUM funciona mejor cuando utilizamos cualquiera de las dos estrategias de expansión explicadas. No obstante, no consideramos relevantes las diferencias. Los resultados corroboran la conveniencia de integrar conocimiento externo procedente de una ontología específica, en este caso UMLS. Estos resultados ponen de manifiesto que, independientemente del algoritmo utilizado, la expansión de términos usando UMLS mejora considerablemente los resultados.

En el futuro se intentarán aplicar estas técnicas de expansión con UMLS a otros corpus biomédicos para comprobar su rendimiento. Por otro lado, se tiene pensado aplicar las mismas estrategias seguidas en este trabajo sobre otras tareas de PLN como minería de textos o recuperación de información biomédica.

Bibliografía

2007. CMC. The Computational Medicine Center's 2007 Medical Natural Language Processing Challenge.

Bauer, Eric y Ron Kohavi. 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, 36(1-2):105–13, August.

Bodenreider, Olivier. 2004. The Unified Medical Language System (UMLS): inte-

grating biomedical terminology. *Nucleic Acids Research*, 32.

Bontempi, Gianluca. 2007. A Blocking Strategy to Improve Gene Selection for Classification of Gene Expression Data. *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, 4(2):293–300.

Ceusters, W., F. Buekens, G. De Moor, y A. Waagmeister. 1997. The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition. En *IMIA Working Group 6*, Jacksonville, Florida.

Díaz Galiano, M.C., M.A. García Cumbreras, M.T. Martín Valdivia, A. Montejo Ráez, y L.A. Ureña López. 2007. Using Information Gain to Improve the ImageCLEF 2006 Collection. En *CLEF*, volumen 4730 de *Lecture Notes in Computer Science*, páginas 711–714. Springer.

Friedman, Nir, Dan Geiger, y Moises Goldszmidt. 1997. Bayesian Network Classifiers. *Mach. Learn.*, 29(2-3):131–163.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. *Proceedings of ECML-98, 10th European Conference on Machine Learning. Springer Verlag*, (1398):137–142.

Lewis, David D., Robert E. Schapire, James P. Callan, y Ron Papka. 1996. Training algorithms for linear text classifiers. En Hans-Peter Frei Donna Harman Peter Schäuble, y Ross Wilkinson, editores, *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, páginas 298–306, Zürich, CH. ACM Press, New York, US.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor, y J. Kandola. 2002. The Perceptron Algorithm with Uneven Margins. En *Proceedings of the International Conference of Machine Learning (ICML'2002)*.

Martín Valdivia, M.T., M. García Vega, y L.A. Ureña López. 2003. LVQ for Text Categorization using Multilingual Linguistic Resource. *Neurocomputing*, 55:665'–679.

Martín Valdivia, M.T., A. Montejo Ráez, M.C. Díaz Galiano, y L.A. Ureña López. 2007. Integración de conocimiento en un dominio específico para la categorización multietiqueta. *Procesamiento del Lenguaje Natural*, 38.

Martín Valdivia, M.T., L.A. Ureña López, y M. García Vega. 2007. The learning vector quantization algorithm applied to automatic text classification tasks. *Neural Networks*, 20(6):748–756.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., y Miller, K. 1993. Introduction to WordNet: An On-line Lexical Database.

Mitkov, Ruslan, editor. 2003. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Montejo Ráez, A. y R. Steinberger. 2004. Why keywording matters. *High Energy Physics Libraries Webzine*, (Issue 10), December.

Pouliquen, Bruno, Denis Delamarre, y Pierre Le Beux. 2002. Indexation de textes médicaux par extraction de concepts, et ses utilisations. En A. Morin & P. Sébillot (eds.), editor, *6th International Conference on the Statistical Analysis of Textual Data, JADT'2002*, volumen 2, páginas 617–628, March.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Vieduts-Stokolo, Natasha. 1987. Concept recognition in an automatic text-processing system for the life sciences.

Wright, Lawrence W., Holly K. Grossetta Nardini, Alan R. Aronson, y Thomas C. Rindflesch. 1999. Hierarchical concept indexing of full-text documents in the Unified Medical Language System® Information Sources Map. *Journal of the American Society for Information Science*, 50(6):514–523.